(19) (11) Publication number: **200**

Generated Document.

PATENT ABSTRACTS OF JAPAN

(21) Application number: **2003085671**

(22) Application date: **08.12.00**

(51) Intl. Cl.: **G06F 3/06**

(30) Priority:

(43) Date of application publication: **24.10.03**

(84) Designated contracting states:

(71) Applicant: **TOSHIBA CORP**

(72) Inventor: **SASAMOTO KYOICHI**
TAKAKUWA MASAYU

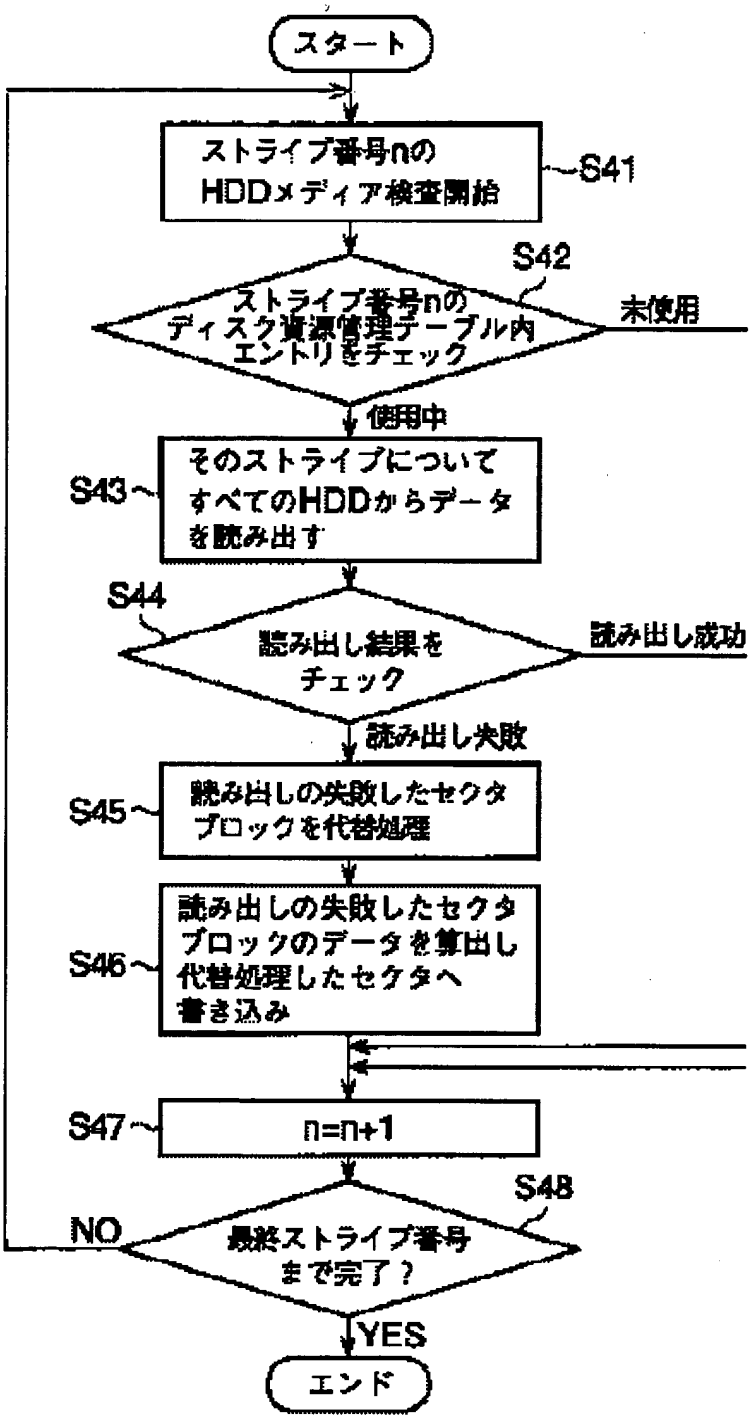
(74) Representative:

(54) METHOD FOR DATA RECOVERY AND DISK ARRAY CONTROLLER IN DISK ARRAY APPARATUS

(57) Abstract:

PROBLEM TO BE SOLVED: To achieve quick detection and recovery of media failures by executing media check processing with a distinction between areas actually in use and not in use at a file system among disk areas of a disk array.

SOLUTION: In a media check processing for checking partial failures of a plurality of HDDs (hard disk drives) composed of the disk array, it is determined whether each stripe in the disk areas of the disk array is used by the file system or not based on a disk source management table (S41, S42), and a media check including data reading from the HDDs is executed only for a stripe (a first stripe) used by the file system (S43, S44).



(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号
特開2003－303057
（P2003－303057A）
(43)公開日　平成15年10月24日(2003.10.24)

(51)Int.Cl.⁷	識別記号	ＦＩ	ﾃｰﾏｺｰﾄﾞ(参考)
Ｇ０６Ｆ　3/06	３０６ ５４０	Ｇ０６Ｆ　3/06	３０６Ｋ　５Ｂ０６５ ５４０

審査請求　未請求　請求項の数４　ＯＬ　（全 16 頁）

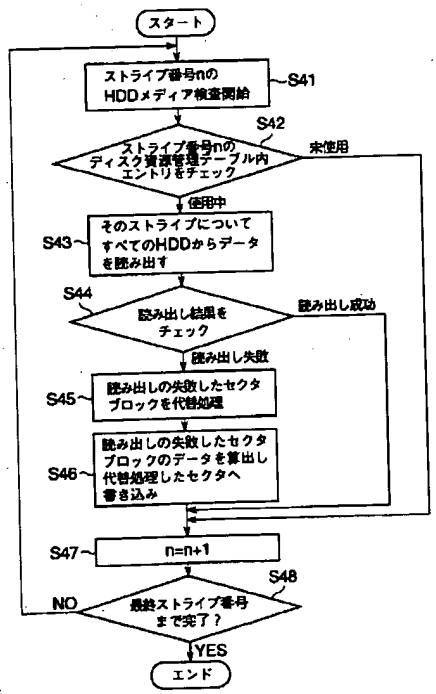
(21)出願番号　特願2003－85671(P2003－85671)
(62)分割の表示　特願2000－374616(P2000－374616)の分割
(22)出願日　平成12年12月８日(2000.12.8)

(71)出願人　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(72)発明者　笹本　享一
東京都府中市東芝町１番地　株式会社東芝府中事業所内
(72)発明者　高桑　正幸
東京都府中市東芝町１番地　株式会社東芝府中事業所内
(74)代理人　100058479
弁理士　鈴江　武彦　（外５名）

Ｆターム(参考)　5B065 BA01 CA30 EA18

(54)【発明の名称】　ディスクアレイ装置におけるデータ修復方法及びディスクアレイコントローラ

(57)【要約】
【課題】ディスクアレイのディスク領域のうちファイルシステムにて実際に使用されている領域と使用されていない領域とを区別してメディア検査処理を実行することで、メディア障害の早期検出・早期修復を実現する。
【解決手段】ディスクアレイを構成する複数のＨＤＤの部分的な障害を検出するメディア検査処理において、ディスクアレイのディスク領域の各ストライプについて、ディスク資源管理テーブルに基づいて、そのストライプがファイルシステムにより使用されているか否かを判定し（Ｓ４１，Ｓ４２）、ファイルシステムにより使用されているストライプ（第１のストライプ）についてのみ、ディスクドライブからのデータ読み出しを含むメディア検査を実行する（Ｓ４３，Ｓ４４）



【特許請求の範囲】

【請求項１】　複数のディスクドライブから構成されるＲＡＩＤ（Redundant Arrays of Inexpensive Disks）構成のディスクアレイを備えたディスクアレイ装置におけるデータ修復方法であって、

前記ディスクアレイのディスク領域を論理ブロック単位に管理するファイルシステムを備えたホスト計算機から、前記ファイルシステムにより使用されている論理ブロックまたは当該論理ブロックを含むストライプを示す第１のディスク資源管理情報を取得するステップと、

前記第１のディスク資源管理情報から前記ディスクアレイのディスク領域内の各ストライプ毎に前記ファイルシステムにより使用されている論理ブロックを含むか否かを示す第２のディスク資源管理情報を生成して前記ディスクアレイ装置内に保持するステップと、

前記ホスト計算機からデータ書き込み要求を受け取った場合、当該要求で指定されたデータ書き込み先が属するストライプを特定するステップと、

前記特定されたストライプが前記ファイルシステムにより使用されている論理ブロックを含む第１のストライプであるか或いは前記ファイルシステムにより使用されている論理ブロックを含まない第２のストライプであるかを、前記ディスクアレイ装置内に保持されている前記第２のディスク資源管理情報に基づいて判定するステップと、

前記特定されたストライプが前記第２のストライプであると判定された場合、当該特定されたストライプが前記第１のストライプであることを示すように前記ディスクアレイ装置内に保持されている前記第２のディスク資源管理情報を更新するステップと、

前記ディスクアレイ内の前記各ディスクドライブの記憶内容を読み出すことにより当該ディスクドライブの部分的な障害を検出するメディア検査処理を実行する場合に、前記ディスクアレイのディスク領域のすべてのストライプについて、前記ファイルシステムにより使用されている論理ブロックを含む第１のストライプであるか、或いは前記ファイルシステムにより使用されている論理ブロックを含まない第２のストライプであるかを、前記第２のディスク資源管理情報に基づいて判定するステップと、

前記判定ステップでの判定結果をもとに、前記第１のストライプについてのみ、前記ディスクドライブからのデータ読み出しを含むメディア検査を実行するステップと、

前記メディア検査で障害が検出された箇所のデータをＲＡＩＤ機能により修復するステップとを具備することを特徴とするディスクアレイ装置におけるデータ修復方法。

【請求項２】　複数のディスクドライブから構成されるＲＡＩＤ（Redundant Arrays of Inexpensive Disks）構成のディスクアレイを備えたディスクアレイ装置におけるデータ修復方法であって、

前記ディスクアレイのディスク領域を論理ブロック単位に管理するファイルシステムを備えたホスト計算機から、前記ファイルシステムにより使用されている論理ブロックまたは当該論理ブロックを含むストライプを示す第１のディスク資源管理情報を取得するステップと、

前記第１のディスク資源管理情報から前記ディスクアレイのディスク領域内の各ストライプ毎に前記ファイルシステムにより使用されている論理ブロックを含むか否かを示す第２のディスク資源管理情報を生成して前記ディスクアレイ装置内に保持するステップと、

前記ホスト計算機からデータ書き込み要求を受け取った場合、当該要求で指定されたデータ書き込み先が属するストライプを特定するステップと、

前記特定されたストライプが前記ファイルシステムにより使用されている論理ブロックを含む第１のストライプであるか或いは前記ファイルシステムにより使用されている論理ブロックを含まない第２のストライプであるかを、前記ディスクアレイ装置内に保持されている前記第２のディスク資源管理情報に基づいて判定するステップと、

前記特定されたストライプが前記第２のストライプであると判定された場合、当該特定されたストライプが前記第１のストライプであることを示すように前記ディスクアレイ装置内に保持されている前記第２のディスク資源管理情報を更新するステップと、

前記ディスクアレイ内の前記各ディスクドライブの記憶内容を読み出すことにより当該ディスクドライブの部分的な障害を検出するメディア検査処理を前記ディスクアレイのディスク領域内の全てのストライプについてストライプ毎に順次実行するステップと、

前記メディア検査処理で前記ディスクドライブの部分的な障害が検出された場合、当該障害が検出された箇所を含むストライプが前記ファイルシステムにより使用されている論理ブロックを含む第１のストライプであるか、或いは前記ファイルシステムにより使用されている論理ブロックを含まない第２のストライプであるかを、前記ディスクアレイ装置内に保持されている前記第２のディスク資源管理情報に基づいて判定するステップと、

前記判定ステップで前記第１のストライプであると判定され、且つ前記ディスクアレイ内の前記複数のディスクドライブのうち前記障害が検出された前記ディスクドライブ以外のディスクドライブが正常な場合、前記障害が検出された箇所のデータをＲＡＩＤ機能により修復するステップと、

前記判定ステップで前記第２のストライプであると判定された場合、前記障害が検出された箇所のデータを固定データにより修復するステップとを具備することを特徴とするディスクアレイ装置におけるデータ修復方法。

【請求項３】 複数のディスクドライブから構成されるＲＡＩＤ（Redundant Arrays of Inexpensive Disks）構成のディスクアレイを制御するディスクアレイコントローラにおいて、
前記ディスクアレイのディスク領域を論理ブロック単位に管理するファイルシステムを備えたホスト計算機から送信される、前記ファイルシステムにより使用されている論理ブロックまたは当該論理ブロックを含むストライプを示す第１のディスク資源管理情報から、前記ディスクアレイのディスク領域内の各ストライプ毎に前記ファイルシステムにより使用されている論理ブロックを含むか否かを示す第２のディスク資源管理情報を生成する手段と、
前記第２のディスク資源管理情報が格納されるメモリと、
前記ホスト計算機からデータ書き込み要求を受け取った場合、当該要求で指定されたデータ書き込み先が属するストライプを特定する手段と、
前記特定手段により前記データ書き込み先が属するストライプが特定された場合に、当該ストライプが、前記ファイルシステムにより使用されている論理ブロックを含む第１のストライプであるか、或いは前記ファイルシステムにより使用されている論理ブロックを含まない第２のストライプであるかを判定する第１の判定手段と、
前記特定手段により特定されたストライプが前記第２のストライプであると前記第１の判定手段により判定された場合、当該ストライプが前記第１のストライプであることを示すように前記メモリに格納されている前記第２のディスク資源管理情報を更新する手段と、
前記ディスクアレイのディスク領域のすべてのストライプについて、前記ファイルシステムにより使用されている論理ブロックを含む第１のストライプであるか、或いは前記ファイルシステムにより使用されている論理ブロックを含まない第２のストライプであるかを、前記第２のディスク資源管理情報に基づいて判定する第２の判定手段と、
前記ディスクアレイ内の前記各ディスクドライブの記憶内容を読み出すことにより当該ディスクドライブの部分的な障害を検出するメディア検査手段であって、前記第２の判定手段により前記第１のストライプであると判定されたストライプだけを対象に、前記ディスクドライブからのデータ読み出しを含むメディア検査を実行するメディア検査手段と、
前記メディア検査手段により障害が検出された箇所のデータをＲＡＩＤ機能により修復するデータ修復手段とを具備することを特徴とするディスクアレイコントローラ。
【請求項４】 複数のディスクドライブから構成されるＲＡＩＤ（Redundant Arrays of Inexpensive Disks）構成のディスクアレイを制御するディスクアレイコントローラにおいて、
前記ディスクアレイのディスク領域を論理ブロック単位に管理するファイルシステムを備えたホスト計算機から送信される、前記ファイルシステムにより使用されている論理ブロックまたは当該論理ブロックを含むストライプを示す第１のディスク資源管理情報から、前記ディスクアレイのディスク領域内の各ストライプ毎に前記ファイルシステムにより使用されている論理ブロックを含むか否かを示す第２のディスク資源管理情報を生成する手段と、
前記第２のディスク資源管理情報が格納されるメモリと、
前記ホスト計算機からデータ書き込み要求を受け取った場合、当該要求で指定されたデータ書き込み先が属するストライプを特定する手段と、
前記特定手段により前記データ書き込み先が属するストライプが特定された場合に、当該ストライプが、前記ファイルシステムにより使用されている論理ブロックを含む第１のストライプであるか、或いは前記ファイルシステムにより使用されている論理ブロックを含まない第２のストライプであるかを判定する第１の判定手段と、
前記特定手段により特定されたストライプが前記第２のストライプであると前記第１の判定手段により判定された場合、当該ストライプが前記第１のストライプであることを示すように前記メモリに格納されている前記第２のディスク資源管理情報を更新する手段と、
前記ディスクアレイ内の前記各ディスクドライブの記憶内容を読み出すことにより当該ディスクドライブの部分的な障害を検出するメディア検査処理を前記ストライプ毎に実行するメディア検査手段と、
前記メディア検査手段により前記ディスクドライブの部分的な障害が検出された場合、当該障害が検出された箇所を含むストライプが前記ファイルシステムにより使用されている論理ブロックを含む第１のストライプであるか、或いは前記ファイルシステムにより使用されている論理ブロックを含まない第２のストライプであるかを、前記第２のディスク資源管理情報に基づいて判定する第２の判定手段と、
前記第２の判定手段により前記第１のストライプであると判定され、且つ前記ディスクアレイ内の前記複数のディスクドライブのうち前記障害が検出された前記ディスクドライブ以外のディスクドライブが正常な場合、前記障害が検出された箇所のデータをＲＡＩＤ機能により修復する第１のデータ修復手段と、
前記第２の判定手段により前記第２のストライプであると判定された場合、前記障害が検出された箇所のデータを固定データにより修復する第２のデータ修復手段とを具備することを特徴とするディスクアレイコントローラ。
【発明の詳細な説明】

【0001】
【発明の属する技術分野】本発明は、ＲＡＩＤ（Redundant Arrays of Inexpensive Disks）構成のディスクアレイ装置におけるデータ修復方法に係り、特にディスクアレイを構成するメンバーのディスクドライブの部分的な障害（メディア障害）を検出して修復する場合に好適なデータ修復方法及びディスクアレイコントローラに関する。
【0002】
【従来の技術】一般にディスクアレイ装置は、複数のディスクドライブ、問えば磁気ディスク装置（以下、ＨＤＤと称する）から構成されるディスクアレイと、このディスクアレイ内の各ＨＤＤ（メンバーＨＤＤ）に対するアクセスを制御するディスクアレイコントローラとを備え、当該各ＨＤＤを並列に動かして読み出し／書き込みを分散して実行することでアクセスの高速化を図ると共に、冗長構成によって信頼性の向上を図るようにした外部記憶装置として知られている。
【0003】上記ディスクアレイコントローラは、ホスト計算機から転送される書き込みデータに対して、データ訂正情報としての冗長データを生成し、上記複数のＨＤＤのうちのいずれかに書き込むようになっている。これにより、複数のＨＤＤのうちの１台の故障に対し、この冗長データと残りのＨＤＤのデータを用いて故障したＨＤＤのデータを修復することを可能としている。
【0004】データ冗長化の手法の１つとして、ＲＡＩＤの手法が知られている。ＲＡＩＤ手法では、ＲＡＩＤのデータと冗長データとの関連において、種々のＲＡＩＤレベルに分類されている。ＲＡＩＤレベルの代表的なものにレベル３とレベル５がある。
【0005】レベル３（ＲＡＩＤレベル３）では、ホスト計算機から転送される更新データ（書き込みデータ）を分割して、その分割された更新データ間の排他的論理和演算を行うことで冗長データとしてのパリティデータを生成し、当該パリティデータで複数のＨＤＤのいずれかに書き込まれている元のパリティデータを更新する。一方、レベル５（ＲＡＩＤレベル５）では、ホスト計算機から転送される更新データ（新データ）と、当該更新データの格納先となるＨＤＤ内領域に格納されている更新前のデータ（旧データ）と、当該更新データの格納先に対応する別のＨＤＤの領域に格納されている更新前のパリティデータ（旧パリティデータ）との間の排他的論理和演算を行うことで、更新されたパリティデータ（新パリティデータ）を生成し、当該更新パリティデータで元のパリティデータを更新する。
【0006】このようなＲＡＩＤ構成のディスクアレイ装置では、ディスクアレイ内のメンバーＨＤＤが故障した場合に、故障したＨＤＤ以外のＨＤＤから、ディスクアレイのディスク領域を管理する単位であるストライプ毎にデータを読み出して、それらのデータの排他的論理和演算を行うＲＡＩＤの機能により、故障したＨＤＤのすべての領域のデータを、故障したＨＤＤに代えて用いられるＨＤＤ内に修復することができる。この故障したＨＤＤに代えて用いられるＨＤＤは、故障したＨＤＤと交換して用いられるＨＤＤ、またはディスクコントローラに予め接続されていて、故障したＨＤＤの代替として割り付けられるスペアＨＤＤである。
【0007】
【発明が解決しようとする課題】このように、ＲＡＩＤ構成のディスクアレイ装置では、ＨＤＤが故障しても、故障したＨＤＤのデータを元通りに修復することができる。
【0008】ところが従来のディスクアレイ装置では、故障したＨＤＤ内のデータを元通りに修復するのに、すべてのＨＤＤ領域のデータをＲＡＩＤの機能により修復していた。このため、近年のようにＨＤＤ容量が増加するのに伴い、データの修復に非常に時間がかかるという問題があった。
【0009】また、データ修復をしている最中は、一般にＲＡＩＤによるデータの冗長性が損なわれる。このため、修復に時間がかかるほど更に他のＨＤＤ故障も発生しやすくなって、データ修復不能となり、データが消失する危険性が高まる。また、故障したＨＤＤのデータを修復するためには、その他のＨＤＤの全領域を読み出す必要がある。もし、この読み出しの対象となるＨＤＤにてメディア障害（ＨＤＤの部分的な障害）が発生すると、ＨＤＤの多重障害となってデータ修復不能となるため、ディスクアレイのディスク領域（ディスクボリューム）の閉塞またはデータ修復処理の継続不能となる。これによりディスクアレイ装置の信頼性が低下する。
【0010】ところが本発明者は、ディスクアレイ装置を利用するホスト計算機のファイルシステムが実際に使用しているＨＤＤ領域以外は、データ修復処理は必ずしも必要でないことを想到するに至った。そこで、ファイルシステムが実際に使用しているＨＤＤ領域（ディスク領域）のみを対象にＨＤＤ（ディスクドライブ）のデータ修復を行うならば、データ修復処理に要する時間を短縮して、ＨＤＤ故障などの危険性を減らすことが可能となる。しかし、従来のディスクアレイ装置は、ホスト計算機のファイルシステムが実際に使用しているＨＤＤ領域を知ることができず、したがってファイルシステムが使用しているＨＤＤ領域のみを対象にＨＤＤのデータ修復を行うことはできない。
【0011】また、ＨＤＤのメディア障害、つまりＨＤＤの部分的な障害であるセクタブロックの障害を早期に検出しこれを修復する目的で、周期的にＨＤＤの内容を読み出して検査するＨＤＤメディア検査処理が一般に知られている。しかし従来のディスクアレイ装置では、ファイルシステムが使用しているＨＤＤ領域が不明のため、ＨＤＤの全領域について読み出し検査をする必要が

あった。このため、故障したＨＤＤのデータの修復処理の場合と同様に、検査に非常に時間を要するという問題があった。

【００１２】また、メディア障害を発見した際には、その障害箇所のセクタブロックを他のセクタブロックに代替する処理を行い、その代替先のブロック内に、代替元のデータを修復する必要がある。しかし、このデータ修復が行えるのは、ＲＡＩＤの機能によりデータの冗長性が確保されている場合に限られる。したがって、ＨＤＤ故障などに起因して行われるデータの修復中は、ＨＤＤメディア検査処理にてメディア障害を見つけても修復することができず、ディスクアレイ装置の信頼性が低下するという問題もあった。

【００１３】本発明は上記事情を考慮してなされたものでその目的は、ディスクアレイのディスク領域のうちファイルシステムにて実際に使用されている領域と使用されていない領域とを区別してメディア検査処理を実行することで、メディア障害の早期検出・早期修復を実現し、これによりディスクアレイ装置の信頼性を向上できるようにすることにある。

【００１４】

【課題を解決するための手段】本発明は、ディスクアレイのディスク領域を論理ブロック単位に管理するファイルシステムを備えたホスト計算機から、当該ファイルシステムにより使用されている論理ブロックまたは当該論理ブロックを含むストライプを示す第１のディスク資源管理情報を取得するステップと、上記第１のディスク資源管理情報から上記ディスク領域内の各ストライプ毎にファイルシステムにより使用されている論理ブロックを含むか否かを示す第２のディスク資源管理情報を生成してディスクアレイ装置内に保持するステップと、上記ホスト計算機からデータ書き込み要求を受け取った場合、当該要求で指定されたデータ書き込み先が属するストライプを特定するステップと、この特定されたストライプがファイルシステムにより使用されている論理ブロックを含む第１のストライプであるか或いはファイルシステムにより使用されている論理ブロックを含まない第２のストライプであるかを、上記ディスクアレイ装置内に保持されている第２のディスク資源管理情報に基づいて判定するステップと、上記特定されたストライプが上記第２のストライプであると判定された場合、当該特定されたストライプが上記第１のストライプであることを示すように上記第２のディスク資源管理情報を更新するステップと、ディスクアレイ内の各ディスクドライブの記憶内容を読み出すことにより当該ディスクドライブの部分的な障害を検出するメディア検査処理を実行する場合に、ディスクアレイのディスク領域のすべてのストライプについて上記第１のストライプであるか、或いは上記第２のストライプであるかを上記第２のディスク資源管理情報に基づいて判定するステップと、第１のストライプ、即ちファイルシステムが使用しているストライプについてのみ、ディスクドライブからのデータ読み出しを含むメディア検査を実行するステップと、このメディア検査で障害が検出された箇所のデータをＲＡＩＤ機能により修復するステップとを備えたことを特徴とする。

【００１５】このような構成においては、ディスクアレイ装置内の例えばディスクアレイコントローラにて、ホスト計算機から第１のディスク資源管理情報が取得され、当該情報から第２のディスク資源管理情報が生成されてディスクアレイ装置内の例えばディスクアレイコントローラに保持される。ここでのメディア検査処理は、ファイルシステムが実際に使用している論理ブロックを含んでいることが第２のディスク資源管理情報によって示されているストライプ（第２のストライプ）だけを対象に実行される。

【００１６】このように、ファイルシステムにて使用されているストライプのみ抽出してメディア検査をすることにより、その検査に要する処理時間を短縮することができる。また、処理時間を短縮したことで、その結果としてメディア障害をより早期に見つけることが可能となり、ディスクドライブの信頼性向上が図れる。これによりディスクアレイ装置の信頼性も向上する。

【００１７】ホスト計算機から取得した第１のディスク資源管理情報は、当該情報を取得した直前までの論理ブロックの使用状況を表す。したがって、それ以降に発生したファイルの更新（ディスクアレイ装置へのデータ書き込み）については反映されていない。しかし、ファイルの更新の度に最新の第１のディスク資源管理情報をホスト計算機から取得するのは大幅な性能低下となる。

【００１８】そこで、ホスト計算機からのデータ書き込み要求をディスクアレイ装置が受け取った場合、つまりディスクアレイ装置でのファイル更新が発生する場合、当該データ書き込み要求に基づいて、ディスクアレイ装置にて保持している第２のディスク資源管理情報を、ストライプの最新の使用状況を表すようにディスクアレイ装置自身が更新するとよい。

【００１９】また本発明は、上記メディア検査処理を全てのストライプについてストライプ毎に順次実行し、ディスクドライブの部分的な障害が検出された場合には、障害が検出された箇所を含むストライプが第１または第２のストライプのいずれであるかを判定し、第１のストライプの場合で、即ちファイルシステムが使用しているストライプである場合で、障害が検出されたディスクドライブ以外のディスクドライブが正常な場合には、障害が検出された箇所のデータをＲＡＩＤ機能により修復し、第２のストライプの場合で、即ちファイルシステムが使用していないストライプである場合には、障害が検出された箇所のデータを固定データにより修復することをも特徴とする。

【００２０】このように、メディア検査で障害が検出さ

れた箇所を含むストライプがファイルシステムにより使用されている場合で、且つ障害が検出されたディスクドライブ以外のディスクドライブが正常な場合だけ、障害が検出された箇所のデータをＲＡＩＤ機能により修復する。これに対し、障害が検出された箇所を含むストライプがファイルシステムにより使用されていない場合には、そのストライプ内のデータを保持している必要はないため、障害が検出されたディスクドライブ以外のディスクドライブが正常であるか否かに無関係に、つまりＲＡＩＤ機能によるデータの冗長性が確保されているか否かに無関係に、障害が検出された箇所のデータを強制的に固定データにより修復することで、メディア障害の修復できる可能性が大幅に向上してディスクドライブの信頼性を向上し、これによりディスクアレイ装置の信頼性も向上する。

【００２１】なお、以上の方法に係る発明は、装置（ディスクアレイコントローラ、または同ディスクアレイコントローラを備えたディスクアレイ装置、または同ディスクアレイ装置を備えた計算機システム）に係る発明としても成立する。

【００２２】

【発明の実施の形態】以下、本発明の実施の形態につき図面を参照して説明する。図１は本発明の一実施形態に係るディスクアレイ装置を備えた計算機システムの構成を示すブロック図である。図１の計算機システムは、ホスト計算機１０と、このホスト計算機１０によって利用されるディスクアレイ装置２０とから構成される。

【００２３】ホスト計算機１０は、当該ホスト計算機１０と接続されているディスクアレイ装置２０（のディスク領域）に格納されるファイルを管理するファイルシステム１１を備えている。このファイルシステム１１はＯＳ（オペレーティングシステム）により提供される機能の一部である。

【００２４】ホスト計算機１０は、ディスクアレイ装置２０のディスク領域内のすべての論理ブロック（連続する複数の物理セクタブロックで構成される固定長のブロック）について、そのブロックがファイルシステム１１により使用されている（つまり有効なデータが格納されている）か、或いは使用されていない（つまりデータは格納されておらず新しいデータを格納できる）かを示す管理テーブル（以下、ディスク資源管理テーブルと称する）１２を、当該計算機１０が持つ記憶装置、例えばＨＤＤ（図示せず）に保持している。この記憶装置がディスクアレイ装置２０であっても構わない。また、ディスク資源管理テーブル１２が、ディスク領域内のすべての物理セクタブロックについて、そのブロックがファイルシステム１１により使用されている否かを示すものであっても構わない。

【００２５】ホスト計算機１０（が持つ記憶装置）には、ファイルシステム１１から予め定められたタイミングでディスク資源管理テーブル１２を取得して、当該テーブル１２から生成されるディスク資源管理情報リスト９０（図９参照）をディスクアレイ装置２０に送信する専用ソフトウエア１３がインストールされている。このディスク資源管理情報リスト９０は、後述するようにファイルシステム１１により使用されているすべての論理ブロックについて、そのブロックの識別情報としての論理ブロック番号の集合からなる。

【００２６】ディスクアレイ装置２０は、ディスクアレイ２１とディスクアレイコントローラ２２とから構成されている。ディスクアレイ２１は、ディスクアレイコントローラ２２と接続される複数のディスクドライブ、例えば４台のＨＤＤ（磁気ディスク装置）２１０-0～２１０-3から構成される。ディスクアレイコントローラ２２には、ＨＤＤ２１０-0～２１０-3のいずれかに障害が発生した場合のバックアップディスク用に割り当てられるスペアＨＤＤ（図示せず）も接続されている。

【００２７】ディスクアレイコントローラ２２は、ディスクアレイ２１内の各ＨＤＤ２１０-0～２１０-3に対するアクセスを制御する。ディスクアレイコントローラ２２は、当該コントローラ２２の主制御部をなすマイクロプロセッサ２２１と、メモリ２２２とを備えている。メモリ２２２には、マイクロプロセッサ２２１が実行する制御プログラム２２２ａが格納されている。またメモリ２２２には、ディスク資源管理テーブル領域２２２ｂが確保されている。このディスク資源管理テーブル領域２２２ｂは、ホスト計算機１０から送信されるディスク資源管理情報リスト９０をもとに生成されるディスク資源管理テーブル１２０を格納するのに用いられる。

【００２８】本実施形態では、ディスクアレイ装置２０がＲＡＩＤ５レベルで用いられるものとする。この場合、ＨＤＤ２１０-0～２１０-3がいずれもデータ並びにパリティデータ（冗長データ）の格納用（データ・パリティディスク用）に用いられる。なお、ディスクアレイ装置２０がＲＡＩＤ３レベルで用いられる場合には、ＨＤＤ２１０-0～２１０-3のうちの３台がデータ格納用（データディスク用）に、残りの１台がパリティデータ格納用（パリティディスク用）に割り当てられる。

【００２９】ディスクアレイ装置２０（内のディスクアレイコントローラ２２）では、ＨＤＤ２１０-0～２１０-3によって実現されるディスクアレイ２１のディスク領域を、図２に示すように複数のストライプ２３に分割して管理する。このストライプ２３のサイズは、１ＨＤＤ当たり６４Ｋ（キロ）バイト～２５６Ｋバイト程度に設定されるのが一般的である。ストライプ２３は、少なくとも１つの論理ブロック２４から構成される。この論理ブロック２４は、ホスト計算機１０のファイルシステム１１がディスクアレイ装置２０（内のディスクアレイ装置２０）のディスク領域を管理する際の最小単位である。つまり、ディスクアレイ装置２０のディスク領域

は、当該ディスクアレイ装置２０ではストライプ２３を単位に管理され、ホスト計算機１０では論理ブロック２４を単位に管理される。通常、論理ブロックのサイズは１Ｋバイト～８Ｋバイト程度である。論理ブロック２４は、連続する複数の物理セクタブロック２５から構成される。このセクタブロック２５のサイズは５１２バイトであるのが一般的である。

【００３０】図３は、ホスト計算機１０内に保持されるディスク資源管理テーブル１２とディスクアレイコントローラ２２のメモリ２２２内のディスク資源管理情報領域２２２ｂに格納されるディスク資源管理テーブル１２０のデータ構造例を示す。

【００３１】ディスク資源管理テーブル１２の各エントリの並び順で決まるエントリ番号は、そのまま論理ブロック番号を表すようになっている。このテーブル１２の各エントリには、そのエントリに固有の論理ブロック番号で表される論理ブロックがファイルシステム１１により使用されているか否かを示すフラグが設定されている。なお、ディスク資源管理テーブル１２の各エントリに、論理ブロック番号と上記フラグとの対が設定されるものであっても構わない。

【００３２】一方、ディスク資源管理テーブル１２０の各エントリの並び順で決まるエントリ番号は、そのままストライプ番号を表すようになっている。このテーブル１２０の各エントリには、そのエントリに固有のストライプ番号で表されるストライプ（に含まれる論理ブロックの少なくとも１つ）がファイルシステム１１により使用されているか否かを示すフラグが設定されている。なお、ディスク資源管理テーブル１２０の各エントリに、ストライプ番号と上記フラグとの対が設定されるものであっても構わない。

【００３３】次に、図１の構成の計算機システムにおける動作を、（１）ホスト計算機１０からのディスク資源管理情報リスト送信時の処理、（２）ホスト計算機１０からのデータ書き込み要求発行時の処理、（３）ディスクアレイ装置２０におけるデータ修復処理、（４）ディスクアレイ装置２０におけるＨＤＤメディア障害検査処理を例に順に説明する。

【００３４】（１）ホスト計算機１０からのディスク資源管理情報リスト送信時の処理

まず、ホスト計算機１０からのディスク資源管理情報リスト送信時の処理について、図４のフローチャートを参照して説明する。

【００３５】ホスト計算機１０は、当該計算機１０（の記憶装置）にインストールされている専用ソフトウエア１３に従い、予め定められたタイミングで、その時点に当該計算機１０（の記憶装置）に保持されているディスク資源管理テーブル１２をファイルシステム１１から取得する。そしてホスト計算機１０は、このディスク資源管理テーブル１２から図９に示すディスク資源管理情報リスト９０を作成し、当該リスト９０をディスクアレイ装置２０に送信する。このときホスト計算機１０は、ディスク資源管理テーブル１２の内容がディスク資源管理情報リスト９０の送信中に変化するのを防止するために、ファイルの更新が発生しないように配慮することが好ましい。また、ディスク資源管理情報リスト９０は極めて大きなサイズとなる可能性があり、その場合には当該リスト９０の送信に長時間を要する。そこで、ディスク資源管理情報リスト９０の送信がホスト計算機１０の効率に影響を及ぼさないように、送信タイミングとして、ホスト計算機１０の立ち上げ時、或いはホスト計算機１０の負荷が少ない夜間等の一定周期を設定するとよい。

【００３６】さて、ホスト計算機１０からディスクアレイ装置２０に送信されるディスク資源管理情報リスト９０は、図９に示すように、ホスト計算機１０内のファイルシステム１１がディスクアレイ装置２０のディスク領域（ディスクボリューム）を扱う際の論理ブロック２４のサイズ９１と、このディスク領域内のすべての論理ブロック２４のうち、ファイルシステム１１により使用されている論理ブロック２４のブロック番号（論理ブロック番号）９２，９２…の集合とから構成される。このように、ディスク資源管理情報リスト９０内に、ファイルシステム１１により使用されていない論理ブロック２４の情報（論理ブロック番号）が含まれていないのは、当該リスト９０のサイズを小さくすることで、当該リスト９０をホスト計算機１０からディスクアレイ装置２０に送信するのに要する時間を短縮するためである。通常、ディスクアレイ装置２０のディスク領域のうち、ファイルシステム１１によって使用されている領域の占める割合は少ない。このような場合、ファイルシステム１１により使用されていない論理ブロック２４の情報をディスク資源管理情報リスト９０に含まないことは、当該リスト９０の送信時間を短縮するのに特に効果がある。なお、ディスク資源管理テーブル１２自体をホスト計算機１０からディスクアレイ装置２０に送信するようにしても構わない。また、ディスク資源管理情報リスト９０またはディスク資源管理テーブル１２０のうちデータ量の少ない方を送信するようにしても構わない。この場合、リスト９０またはテーブル１２０のいずれの送信であるかを示す情報を付加して送信するとよい。

【００３７】ディスクアレイ装置２０内のディスクアレイコントローラ２２（に設けられたマイクロプロセッサ２２１）は、ホスト計算機１０からディスク資源管理情報リスト９０が送信されると、当該リスト９０を受信する（ステップＳ１）。するとディスクアレイコントローラ２２は、ディスク資源管理情報リスト９０に含まれているすべての論理ブロック番号をもとに、ファイルシステム１１により使用されていない論理ブロックの論理ブロック番号を判別し、ディスクアレイ装置２０のディス

ク領域内のすべての論理ブロックについて、そのブロックを示す論理ブロック番号の例えば昇順に、そのブロックが使用されているか否かを示すエントリが配置された、図3に示すディスク資源管理テーブル１２を復元する（ステップＳ２）。

【００３８】次にディスクアレイコントローラ２２は、ホスト計算機１０により管理される論理ブロック番号とディスクアレイ装置２０により管理されるストライプ番号との対応付けを行う（ステップＳ３）。この対応付けは次のように行われる。

【００３９】まずディスクアレイコントローラ２２は、「ストライプ当たりの論理ブロック数」を、自身が管理している「ストライプのサイズ」と、ホスト計算機１０から送信されたディスク資源管理情報リスト９０に含まれている「論理ブロックのサイズ」９１とから、

「ストライプ当たりの論理ブロック数」＝「ストライプのサイズ」／「論理ブロックのサイズ」

により算出する。

【００４０】次にディスクアレイコントローラ２２は、すべての「論理ブロック番号」について、その「論理ブロック番号」と「ストライプ当たりの論理ブロック数」とから、「論理ブロック番号」の示す論理ブロック２４が含まれているストライプ２３を示す「ストライプ番号」を、

「ストライプ番号」＝｛「論理ブロック番号」／「ストライプ当たりの論理ブロック数」｝の整数部

により算出する。例えば、「ストライプ当たりの論理ブロック数」を「４」とすると、「論理ブロック番号」が「０」～「３」の論理ブロックを含むストライプの「ストライプ番号」はいずれも「０」である。この結果、ホスト計算機１０により管理される論理ブロック番号とディスクアレイ装置２０により管理されるストライプ番号との対応付けが行われたことになる。

【００４１】ディスクアレイコントローラ２２は、論理ブロック番号とストライプ番号との対応付けを行うと、その対応付けの結果と復元されたディスク資源管理テーブル１２とから、ディスクアレイ装置２０内のすべてのストライプについて、そのストライプを示すストライプ番号の例えば昇順に、そのストライプがファイルシステム１１により使用されているか否かを示すエントリが配置された、図3に示すディスク資源管理テーブル１２０を作成する（ステップＳ４）。ここでは、ファイルシステム１１によって使用されている論理ブロックを１つでも含むストライプは、ファイルシステム１１により使用されていると判定されて、対応するエントリに使用中を示すフラグが設定される。これに対し、ファイルシステム１１によって使用されている論理ブロックを含まないストライプは、ファイルシステム１１によって使用されていないと判定されて、対応するエントリに未使用（不使用）を示すフラグが設定される。

【００４２】ディスクアレイコントローラ２２はディスク資源管理テーブル１２０を作成すると、当該テーブル１２０をメモリ２２２内のディスク資源管理テーブル領域２２２ｂに上書きコピーする（ステップＳ５）。

【００４３】上記のように、論理ブロック番号とストライプ番号との対応付けをディスクアレイ装置２０で行う場合、ホスト計算機１０ではディスクアレイ装置２０に固有のストライプのサイズを考慮する必要がない。但し、ディスクアレイ装置２０では、ホスト計算機１０に固有の論理ブロックのサイズを考慮する必要がある。

【００４４】これに対し、図9のディスク資源管理情報リスト９０に代えて、図１０のデータ構造のディスク資源管理情報リスト１００を用いるならば、ディスクアレイ装置２０では、ホスト計算機１０に固有の論理ブロックのサイズを考慮する必要がない。この図１０の構造のディスク資源管理情報リスト１００は、ファイルシステム１１により使用されている論理ブロック２４が含まれるストライプを示すストライプ番号１０１，１０１…の集合から構成される。但し、図１０の構造のディスク資源管理情報リスト１００をホスト計算機１０で用意するには、当該ホスト計算機１０がディスクアレイ装置２０内のディスクアレイコントローラ２２から予めストライプのサイズを取得し、論理ブロック番号とストライプ番号との対応付けを、当該ホスト計算機１０にインストールされている専用ソフトウエア１３に従って実行する必要がある。

【００４５】（２）ホスト計算機１０からのデータ書き込み要求発行時の処理

次に、メモリ２２２内のディスク資源管理テーブル領域２２２ｂに、図3に示すディスク資源管理テーブル１２０が格納されている状態で、ホスト計算機１０からディスクアレイ装置２０内のディスクアレイコントローラ２２に対してデータの書き込み要求が発行された場合の処理について、図5のフローチャートを参照して説明する。

【００４６】まずディスクアレイコントローラ２２（内のマイクロプロセッサ２２１）は、ホスト計算機１０（内のファイルシステム１１）からディスクアレイ装置２０に対してデータ書き込み要求が発行されると、当該要求を受信する（ステップＳ１１）。この要求には、アドレス（開始アドレス）とサイズとが含まれている。

【００４７】次にディスクアレイコントローラ２２は、受信したデータ書き込み要求に含まれている開始アドレスとサイズとから、書き込み対象となるストライプを示すストライプ番号を算出する（ステップＳ１２）。

【００４８】次にディスクアレイコントローラ２２は、ディスク資源管理テーブル領域２２２ｂに格納されているディスク資源管理テーブル１２０内のエントリのうち、ステップＳ１２で算出したストライプ番号で指定されるエントリを参照することにより、当該ストライプ番

号の示す書き込み対象ストライプがファイルシステム１１にて既に使用されているか否かを判定する（ステップＳ１３）。もし、書き込み対象ストライプがそれまで使用されていなかった場合、ディスクアレイコントローラ２２はステップＳ１３で参照したディスク資源管理テーブル１２０内のエントリの内容（フラグの状態）を、未使用から使用中を示すように更新する（ステップＳ１４）。このように、ホスト計算機１０からのデータ書き込み要求で指定されるデータ書き込みにより、当該要求で指定される論理ブロックを含むストライプの状態が未使用から使用中に変化すると判定された場合には、当該ストライプに対応するディスク資源管理テーブル１２０内のエントリの内容が使用中を示すように更新される。

【００４９】ディスクアレイコントローラ２２はステップＳ１４を実行すると、ディスクアレイ２１に対してホスト計算機１０からのデータ書き込み要求で指定されたデータ書き込みを行う（ステップＳ１５）。

【００５０】またディスクアレイコントローラ２２は、ステップＳ１２で算出したストライプ番号の示す書き込み対象ストライプがファイルシステム１１にて既に使用されている場合には（ステップＳ１３）、ステップＳ１４をスキップしてステップＳ１５に進み、ホスト計算機１０からのデータ書き込み要求で指定されたデータ書き込みを行う。

【００５１】（３）ディスクアレイ装置２０におけるデータ修復処理

次に、ディスクアレイ装置２０におけるデータ修復処理について、図6のフローチャートを参照して説明する。

【００５２】今、ディスクアレイ２１内のHDD２１０-0～２１０-3のうちHDD２１０-3が故障したために、その故障したHDD（旧HDD）２１０-3を新たなHDD（新HDD）２１０-3に交換して、旧HDD２１０-3内のデータを新HDD２１０-3に修復するものとする。ここでは便宜的に、新HDD、つまり修復先となるHDDにも、故障した旧HDD２１０-3と同一符号“２１０-3”を付してある。なお、修復先となるHDDがディスクアレイコントローラ２２に予め接続されているスペアHDDであっても構わない。

【００５３】ディスクアレイコントローラ２２（内のマイクロプロセッサ２２１）はHDD２１０-3が故障した場合、その故障したHDD（旧HDD）２１０-3内のデータを、新HDD２１０-3に修復する処理を、ストライプ番号ｎが０の先頭のストライプから順番に次のように実行する（ステップＳ２１）。

【００５４】まずディスクアレイコントローラ２２は、ストライプ番号ｎ（ｎの初期値は０）のストライプ２３のデータ修復のために、そのストライプ番号ｎで指定されるディスク資源管理テーブル１２０内のエントリを参照して、そのストライプ番号ｎの示すストライプ２３（に含まれている論理ブロックの少なくとも１つ）がファイルシステム１１により使用されているか否かを判定する（ステップＳ２２）。

【００５５】もし、ストライプ番号ｎの示すストライプ２３が使用されているならば、ディスクアレイコントローラ２２は、従来から知られているＲＡＩＤの機能に従う通常のデータ修復処理を以下に述べる手順で図１１に示すように行う。

【００５６】まず、ディスクアレイコントローラ２２は、修復するストライプ２３について、正常なすべてのHDD２１０-0～２１０-2からのデータ読み出し１１１を行う（ステップＳ２３）。次にディスクアレイコントローラ２２は、読み出したデータを使用してＲＡＩＤの機能により、つまり排他的論理和演算１１２により、その演算結果として修復されたデータを取得する（ステップＳ２４）。そしてディスクアレイコントローラ２２は、取得したデータをストライプ２３に含まれる新HDD２１０-3内の領域に書き込む動作１１３を実行する（ステップＳ２５）。これにより旧HDD２１０-3内のデータが新HDD２１０-3に修復される。

【００５７】これに対し、ストライプ番号ｎの示すストライプ２３がファイルシステム１１により使用されていないならば、ディスクアレイコントローラ２２は当該ストライプ２３内には修復すべき有効なデータが格納されていないものと判断する。この場合、ディスクアレイコントローラ２２は図１２に示す動作を行う。即ちディスクアレイコントローラ２２は、ストライプ２３に含まれる正常なすべてのHDD２１０-0～２１０-2内の各領域に対して予め定められた固定データを書き込む動作２１１を実行すると共に、その固定データの排他的論理和値をストライプ２３に含まれる新HDD２１０-3内の領域に書き込む動作２１２を実行する（ステップＳ２６）。ここで、固定データの排他的論理和値は、実際に排他的論理和演算を行うことにより取得されるものでも、予め定められた固定値であっても構わない。つまり、ステップＳ２６では、ファイルシステム１１により使用されていないストライプ２３への固定データの書き込みが行われる。このステップＳ２６の動作は、正常なHDD２１０-0～２１０-2からの読み出しを必要としないため、ファイルシステム１１により使用されているストライプ２３のデータを修復する場合（ステップＳ２３～Ｓ２５）に比べて、短時間で実行できる。また、HDD２１０-0～２１０-2からの読み出しが行われないことにより、HDDの多重障害となる危険性が大幅に低下する。

【００５８】ディスクアレイコントローラ２２はステップＳ２５またはＳ２６を終了すると、ストライプ番号ｎを１だけインクリメントし（ステップＳ２７）、そのインクリメント後のストライプ番号ｎが最終ストライプ番号を越えたか否かにより、最終ストライプまで修復処理を終了したか否かを判定する（ステップＳ２８）。ディスクアレイコントローラ２２は、ステップＳ２８で未終

了を判定したならば、ステップＳ２１以降の動作を再度実行し、終了を判定したならば一連のデータ修復処理を終了する。

【００５９】なお、以上に述べた故障したＨＤＤ内のデータを修復する処理では、ファイルシステム１１により使用されていないストライプについては、当該ストライプへの固定データの書き込み（ステップＳ２６）が行われるものとして説明した。しかし、ファイルシステム１１により使用されていないストライプ内には修復すべき有効なデータが格納されていないことから、図６のフローチャートにおいて破線６０で示すように、このステップＳ２６の動作（固定データによるストライプの修復動作）をスキップするようにしても構わない。

【００６０】しかし、ステップＳ２６をスキップすると、ファイルシステム１１により使用されていないストライプについては、ＲＡＩＤレベル５におけるデータと冗長データ（パリティデータ）との整合性が得られなくなる。つまり、データに対して正しいパリティデータが生成されていない状態となる。したがって、ステップＳ２６の動作をスキップする方法を適用する場合、ホスト計算機１０からのデータ書き込み要求を、図７のフローチャートに示す手順に従って図１３に示すように処理する必要がある。

【００６１】まずディスクアレイコントローラ２２は、ホスト計算機１０からディスクアレイ装置２０に対してデータ書き込み要求が発行された場合、当該要求で指定されたデータの書き込み対象となるストライプ２３のストライプ番号を算出し、そのストライプ２３がファイルシステム１１により使用されているか否かをディスク資源管理テーブル１２０に基づいて判定する（ステップＳ３１～Ｓ３３）。ここまでは図５のフローチャートのステップＳ１１～Ｓ１３と同様である。

【００６２】もし、データの書き込み対象となるストライプ２３がファイルシステム１１により使用されていない場合、ディスクアレイコントローラ２２は当該ストライプ２３はＲＡＩＤレベル５によるパリティデータの整合性が得られていないものと判断する。この場合、まずディスクアレイコントローラ２２は、ホスト計算機１０からのデータ書き込み要求で指定された書き込みデータ（新規書き込みデータ）１３１と予め定められた、“ディスクアレイ２１を構成するＨＤＤの数－２”個のＨＤＤ用の固定データ１３３との排他的論理和値１３５を正しいパリティデータ（冗長データ）として取得する（ステップＳ３４）。ここで、固定データ１３３を全ビットが“０”のデータとするならば、排他的論理和値１３５は書き込みデータ１３１に一致する。この場合、書き込みデータ１３１を排他的論理和値（冗長データ）１３５とすることができるため、排他的論理和演算を必要としない。

【００６３】次にディスクアレイコントローラ２２は、データの書き込み対象となるストライプ２３に含まれるすべてのＨＤＤ２１０-0～２１０-3内の領域に、各ＨＤＤ毎に、書き込みデータ１３１、固定データ１３３、または排他的論理和値（冗長データ）１３５を書き込む（ステップＳ３５）。ここでは、ホスト計算機１０からのデータ書き込み要求で指定されたデータの書き込み先ＨＤＤがＨＤＤ２１０-0であり、ストライプ２３において冗長データが格納されているＨＤＤがＨＤＤ２１０-3であるものとすると、ＨＤＤ２１０-0に対する書き込みデータ１３１の書き込み１３２と、ＨＤＤ２１０-1及び２１０-2に対する固定データ１３３の書き込み１３４と、ＨＤＤ２１０-3に対する排他的論理和値（冗長データ）１３５の書き込み１３６とが、それぞれ行われる。これにより、故障したＨＤＤのデータ修復の際に修復をスキップしたストライプについてもデータの冗長性を保証できる。

【００６４】ディスクアレイコントローラ２２はステップＳ３５を終了すると、ステップＳ３３で参照したディスク資源管理テーブル１２０内のエントリの内容、つまりデータの書き込み対象となったストライプ２３の使用の有無を示すディスク資源管理テーブル１２０内のエントリの内容を、未使用から使用中を示すように更新する（ステップＳ３６）。

【００６５】なお、データの書き込み対象となるストライプ２３がファイルシステム１１により使用されている場合には、ディスクアレイコントローラ２２は通常のＲＡＩＤ手法によるデータ書き込みを行う（ステップＳ３７）。ここでは、データ書き込み要求で指定されたデータ（新データ）と、当該新データの格納先となるＨＤＤ内領域に格納されているデータ（旧データ）と、同じストライプ２３に含まれている別のＨＤＤ内領域に格納されているパリティデータ（旧パリティデータ）との間の排他的論理和演算を行うことで、新パリティデータ（新冗長データ）を生成し、当該新パリティデータで旧パリティデータを更新する。

【００６６】（４）ディスクアレイ装置２０におけるＨＤＤメディア障害検査処理

次に、ディスクアレイ装置２０におけるＨＤＤメディア障害検査処理について、図８のフローチャートを参照して説明する。

【００６７】ディスクアレイコントローラ２２は、ＨＤＤメディア検査を例えば当該コントローラ２２の有するパトロール機能により周期的に実行する。ここではディスクアレイコントローラ２２は、ＨＤＤメディア検査を、ストライプ番号ｎが０の先頭のストライプから順番に次のように実行する（ステップＳ４１）。

【００６８】まずディスクアレイコントローラ２２は、ストライプ番号ｎ（ｎの初期値は０）のストライプ２３のＨＤＤメディア検査のために、そのストライプ番号ｎで指定されるディスク資源管理テーブル１２０内のエン

トリを参照して、そのストライプ番号nの示すストライプ23がファイルシステム11により使用されているか否かを判定する（ステップS42）。

【0069】もし、ストライプ番号nの示すストライプ23が使用されているならば、ディスクアレイコントローラ22は、当該ストライプについて、すべてのHDD210-0～210-3からのデータ読み出しを行う（ステップS43）。

【0070】次にディスクアレイコントローラ22は、HDD210-0～210-3からのデータ読み出し結果をチェックして、読み出しに成功したか否かを判定する（ステップS44）。

【0071】もし、HDD210-0～210-3のいずれかからのデータ読み出しに失敗したならば、ディスクアレイコントローラ22は、その失敗したセクタブロック、つまりメディア障害が検出された不良セクタブロックを、同じHDD内の別のセクタ（交代セクタ）に代替する代替処理を行う（ステップS45）。例えば、図14に示すように、HDD210-1内のセクタブロック141が不良セクタブロックとして検出された場合であれば、当該セクタブロック141を同じHDD210-1内の任意の交代セクタ142に代替する代替処理143が行われる。

【0072】次にディスクアレイコントローラ22は、RAIDの機能を使用して、不良セクタブロック141の修復されたデータを算出し、そのデータを交代セクタ142に書き込む動作144を行う（ステップS46）。そしてディスクアレイコントローラ22はステップS47に進む。

【0073】これに対し、ストライプ番号nの示すストライプ23についてHDD210-0～210-3からのデータ読み出しに成功したならば、ディスクアレイコントローラ22はステップS45，S46ををスキップしてステップS47に進む。

【0074】また、ストライプ番号nの示すストライプ23がファイルシステム11によって使用されていないならば、ディスクアレイコントローラ22は当該ストライプ23の検査をせずに、ステップS43～S46をスキップしてステップS47に進む。

【0075】ディスクアレイコントローラ22は、ステップS47においてストライプ番号nを1だけインクリメントし、そのインクリメント後のストライプ番号nが最終ストライプ番号を越えるまで（ステップS48）、ステップS41以降の動作を繰り返す。

【0076】以上に述べたHDDメディア障害検査処理では、ファイルシステム11により使用されているストライプのみ、HDDからのデータ読み出しによる検査（メディア検査）を実行する場合について説明したが、これに限るものではない。例えば図15のフローチャートに示すように、ディスクアレイ装置20のディスク領域のすべてのストライプについて、その使用の有無に無関係にHDDからのデータ読み出しによる検査を実行するようにしてもよい（ステップS52，S53）。ここでは、データ読み出しによる検査で、メディア障害となったセクタブロック（不良セクタブロック）が検出された場合に、ディスク資源管理テーブル120を参照して、そのセクタブロックを含むストライプがファイルシステム11により使用されているか否かを判定する（ステップS54）。

【0077】もし、上記ストライプがファイルシステム11により使用されていない場合、そのストライプ内のデータを保持する必要性はない。そこで、この場合には、RAIDの冗長性の有無に拘わらずに、図12に示すように不良セクタブロック（141）の代替処理（143）を行った後に（ステップS55）、交代セクタ（142）に（修復データではなくて）予め定められた固定データの書き込み（144）を行って修復する（ステップS56）。このステップS56では、上記ステップS26におけるストライプに対するのと同様に、交代セクタに対応する他の正常なHDD内のセクタに固定データが書き込まれる。

【0078】一方、上記ストライプがファイルシステム11により使用されている場合、他のHDDにて故障などが発生しておらず（ステップS57）、したがってデータの冗長性が確保されている場合は、不良セクタブロック（141）を交代セクタ（142）に代替する処理（143）を行った後（ステップS58）、RAIDの機能を使用して、不良セクタブロック（141）のデータを修復し、その修復データの交代セクタ（142）への書き込み（144）を行う（ステップS59）。

【0079】これに対し、他のHDDが故障しているためにデータの冗長性がなくなっている場合は（ステップS57）、不良セクタブロック（141）のデータをRAIDの機能により修復することができない。この場合、上記ステップS58，S59をスキップし、不良セクタブロック（141）をそのまま放置する。

【0080】なお、以上の実施形態では、ディスクアレイ装置20がRAID5レベルで用いられるものとして説明したが、本発明は、RAID3レベルなど、他のRAIDレベルで用いられるディスクアレイ装置にも、データ修復の方法は異なるものの、RAID5レベルの場合と同様に適用できる。

【0081】なお、本発明は、上記実施形態に限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で種々に変形することが可能である。更に、上記実施形態には種々の段階の発明が含まれており、開示される複数の構成要件における適宜な組み合わせにより種々の発明が抽出され得る。例えば、実施形態に示される全構成要件から幾つかの構成要件が削除されても、発明が解決しようとする課題の欄で述べた課題が解決でき、発明

の効果の欄で述べられている効果の少なくとも１つが得られる場合には、この構成要件が削除された構成が発明として抽出され得る。

【００８２】

【発明の効果】以上詳述したように本発明によれば、ディスクアレイのディスク領域のうちファイルシステムにて実際に使用されている領域を判定し、このファイルシステムにて実際に使用されている領域に絞ってディスクドライブのメディア障害検査を行うようにしたので、処理時間を短縮することができ、その結果、メディア障害の早期検出が可能となり、ディスクアレイ装置の信頼性を向上することができる。

【００８３】また本発明によれば、ホスト計算機からデータ書き込み要求を受け取った場合、当該要求で指定されたデータ書き込み先が属するストライプが、ファイルシステムにより使用されている論理ブロックを含む第１のストライプであるか或いはファイルシステムにより使用されている論理ブロックを含まない第２のストライプであるかを、ディスクアレイ装置内に保持されているディスク資源管理情報（第２のディスク資源管理情報）に基づいて判定し、第２のストライプであるならば、当該データ書き込み先が属するストライプが第１のストライプであることを示すように第２のディスク資源管理情報を更新するようにしたので、当該第２のディスク資源管理情報によりストライプの最新の使用状況を表すことができ、ファイルの更新の度に最新の第１のディスク資源管理情報をホスト計算機から取得しなくても済む。

【図面の簡単な説明】

【図１】本発明の一実施形態に係るディスクアレイ装置２０を備えた計算機システムの構成を示すブロック図。

【図２】ディスクアレイ２１のディスク領域を管理するのに用いられるストライプ、論理ブロック及びセクタブロックの関係を説明するための図。

【図３】ホスト計算機１０内に保持されるディスク資源管理テーブル１２とディスクアレイコントローラ２２のメモリ２２２内のディスク資源管理情報領域２２２ｂに格納されるディスク資源管理テーブル１２０のデータ構造例を示す図。

【図４】ディスク資源管理情報リスト送信時の処理手順を示すフローチャート。

【図５】ホスト計算機１０からのデータ書き込み要求発行時の処理手順を示すフローチャート。

【図６】ディスクアレイ装置２０におけるデータ修復処理手順を示すフローチャート。

【図７】ホスト計算機１０からのデータ書き込み要求発行時の処理手順の変形例を示すフローチャート。

【図８】ディスクアレイ装置２０におけるＨＤＤメディア障害検査の処理手順を示すフローチャート。

【図９】ホスト計算機１０からディスクアレイ装置２０に送信されるディスク資源管理情報リストのデータ構造例を示す図。

【図１０】上記ディスク資源管理情報リストの変形例を示す図。

【図１１】ＲＡＩＤレベル５における通常のデータ修復処理を説明するための図。

【図１２】使用されていないストライプへの固定データ書き込み時の動作を説明するための図。

【図１３】使用されていないストライプへの新規データ書き込み時の動作を説明するための図。

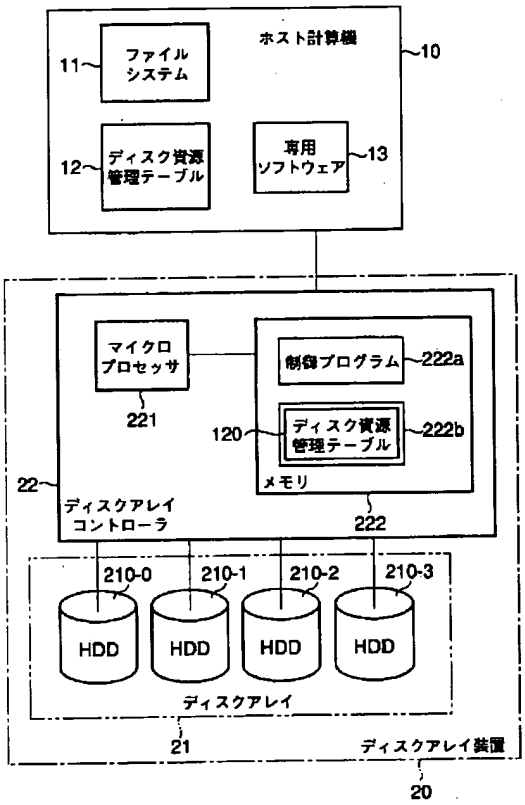
【図１４】ＨＤＤメディア障害検査で不良セクタブロックが検出された場合のデータ修復動作を説明するための図。

【図１５】図８に示したＨＤＤメディア障害検査の処理手順の変形例を示すフローチャート

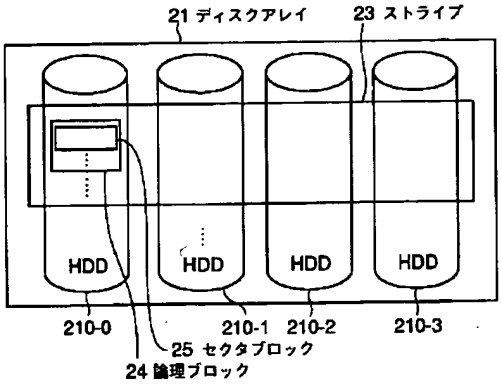
【符号の説明】

１０…ホスト計算機
１１…ファイルシステム
１２…ディスク資源管理テーブル
１３…専用ソフトウエア
２０…ディスクアレイ装置
２１…ディスクアレイ
２２…ディスクアレイコントローラ
２３…ストライプ
２４…論理ブロック
２５…セクタブロック
９０，１００…ディスク資源管理情報リスト（第１のディスク資源管理情報）
１２０…ディスク資源管理テーブル（第２のディスク資源管理情報）
２１０-0～２１０-3…ＨＤＤ（ディスクドライブ）
２２２ｂ…ディスク資源管理テーブル領域

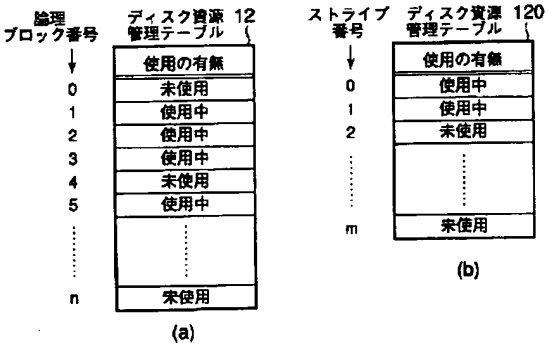
【図1】


ホスト計算機
10
11
ファイル
システム
12
ディスク資源
管理テーブル
専用
ソフトウェア
13
マイクロ
プロセッサ
221
制御プログラム
222a
120
ディスク資源
管理テーブル
222b
メモリ
222
22
ディスクアレイ
コントローラ
210-0
210-1
210-2
210-3
HDD
HDD
HDD
HDD
ディスクアレイ
21
ディスクアレイ装置
20

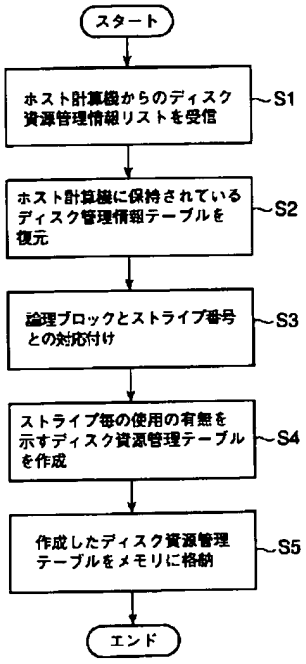
【図2】


21 ディスクアレイ
23 ストライプ
HDD
HDD
HDD
HDD
210-0
210-1
210-2
210-3
25 セクタブロック
24 論理ブロック

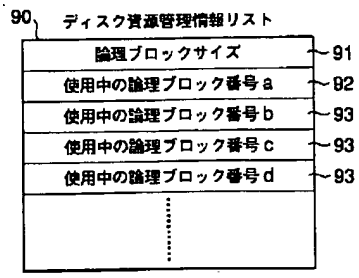
【図3】


論理
ブロック番号
ディスク資源
管理テーブル
12
使用の有無
0
未使用
1
使用中
2
使用中
3
使用中
4
未使用
5
使用中
n
未使用
(a)
ストライプ
番号
ディスク資源
管理テーブル
120
使用の有無
0
使用中
1
使用中
2
未使用
m
未使用
(b)

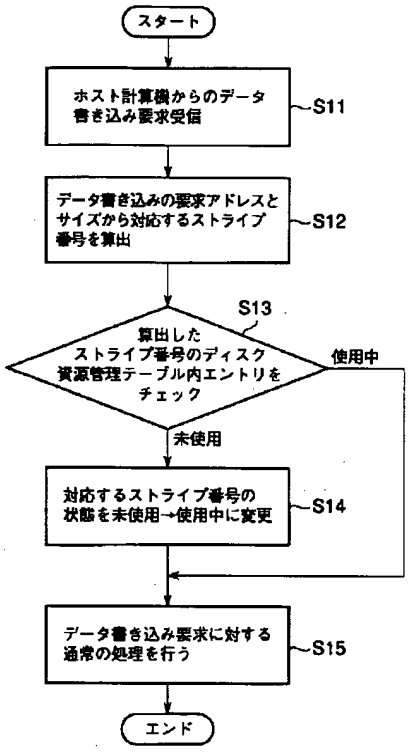
【図4】


スタート
ホスト計算機からのディスク
資源管理情報リストを受信
S1
ホスト計算機に保持されている
ディスク管理情報テーブルを
復元
S2
論理ブロックとストライプ番号
との対応付け
S3
ストライプ毎の使用の有無を
示すディスク資源管理テーブル
を作成
S4
作成したディスク資源管理
テーブルをメモリに格納
S5
エンド

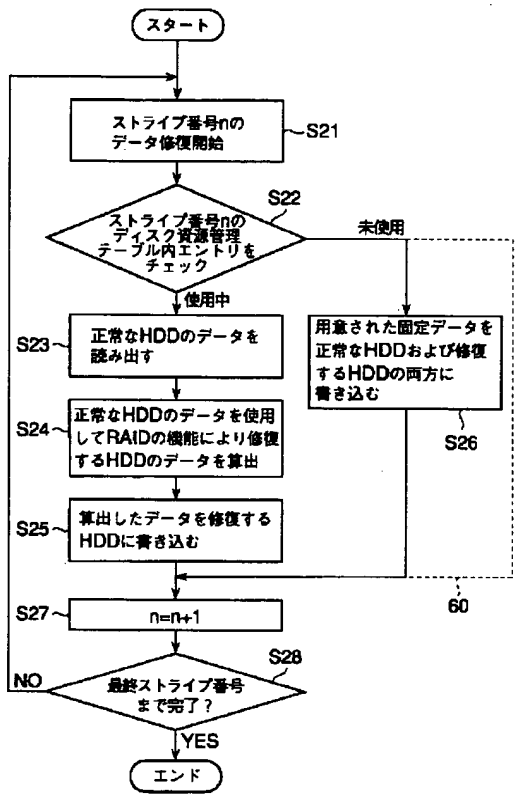
【図9】


90
ディスク資源管理情報リスト
論理ブロックサイズ
91
使用中の論理ブロック番号a
92
使用中の論理ブロック番号b
93
使用中の論理ブロック番号c
93
使用中の論理ブロック番号d
93

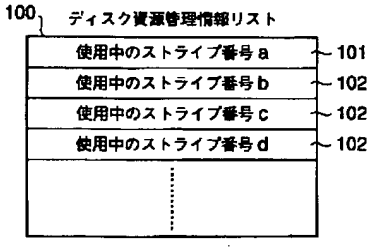
【図5】


スタート
ホスト計算機からのデータ書き込み要求受信
S11
データ書き込みの要求アドレスとサイズから対応するストライプ番号を算出
S12
S13
算出したストライプ番号のディスク資源管理テーブル内エントリをチェック
使用中
未使用
対応するストライプ番号の状態を未使用→使用中に変更
S14
データ書き込み要求に対する通常の処理を行う
S15
エンド

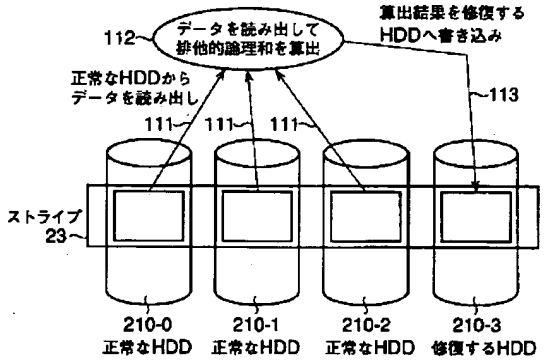
【図6】


スタート
ストライプ番号nのデータ修復開始
S21
S22
ストライプ番号nのディスク資源管理テーブル内エントリをチェック
未使用
使用中
S23
正常なHDDのデータを読み出す
用意された固定データを正常なHDDおよび修復するHDDの両方に書き込む
S26
S24
正常なHDDのデータを使用してRAIDの機能により修復するHDDのデータを算出
S25
算出したデータを修復するHDDに書き込む
60
S27
n=n+1
S28
NO
最終ストライプ番号まで完了？
YES
エンド

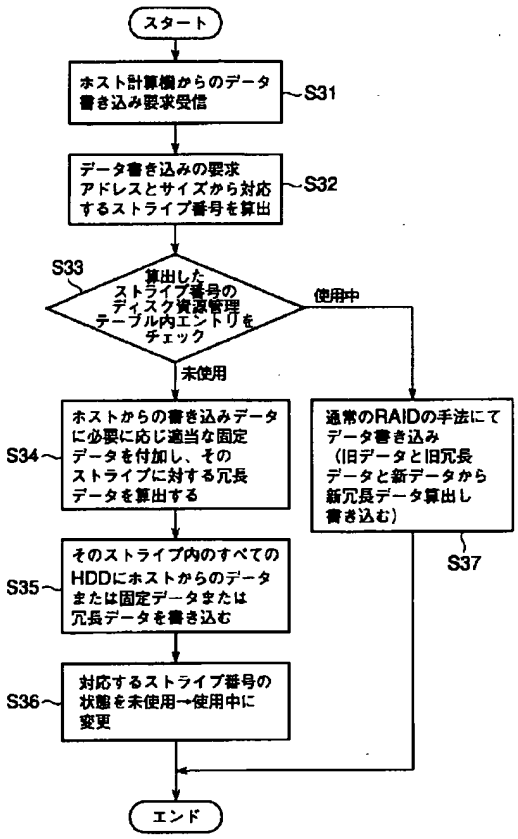
【図10】


100
ディスク資源管理情報リスト
使用中のストライプ番号a
101
使用中のストライプ番号b
102
使用中のストライプ番号c
102
使用中のストライプ番号d
102

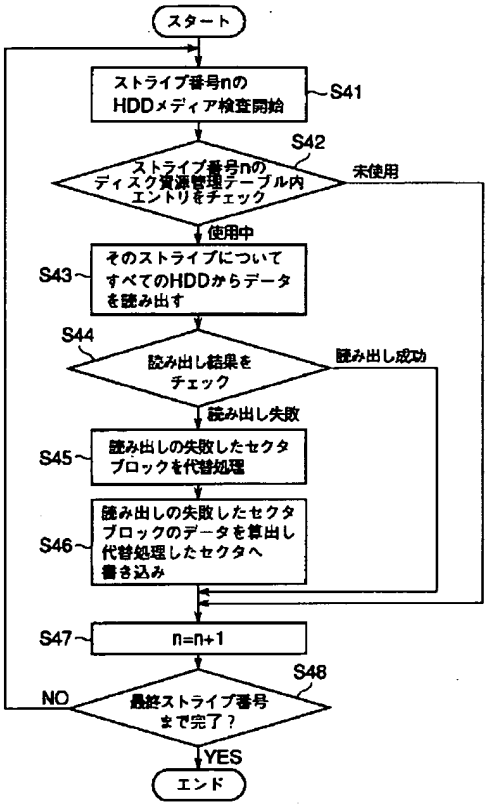
【図11】


112
データを読み出して排他的論理和を算出
算出結果を修復するHDDへ書き込み
113
正常なHDDからデータを読み出し
111
111
111
ストライプ
23
210-0
正常なHDD
210-1
正常なHDD
210-2
正常なHDD
210-3
修復するHDD

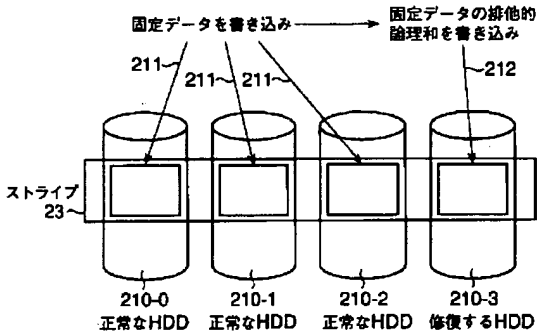
【図7】


スタート
ホスト計算機からのデータ書き込み要求受信
S31
データ書き込みの要求アドレスとサイズから対応するストライプ番号を算出
S32
S33
算出したストライプ番号のディスク資源管理テーブル内エントリをチェック
使用中
未使用
ホストからの書き込みデータに必要に応じ適当な固定データを付加し、そのストライプに対する冗長データを算出する
S34
通常のRAIDの手法にてデータ書き込み（旧データと旧冗長データと新データから新冗長データ算出し書き込む）
S37
そのストライプ内のすべてのHDDにホストからのデータまたは固定データまたは冗長データを書き込む
S35
対応するストライプ番号の状態を未使用→使用中に変更
S36
エンド

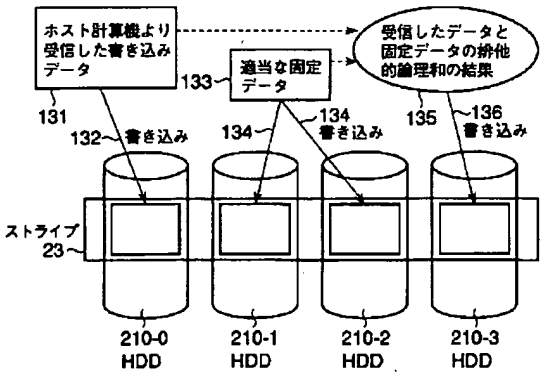
【図8】


スタート
ストライプ番号nのHDDメディア検査開始
S41
S42
ストライプ番号nのディスク資源管理テーブル内エントリをチェック
未使用
使用中
S43
そのストライプについてすべてのHDDからデータを読み出す
S44
読み出し結果をチェック
読み出し成功
読み出し失敗
S45
読み出しの失敗したセクタブロックを代替処理
S46
読み出しの失敗したセクタブロックのデータを算出し代替処理したセクタへ書き込み
S47
n=n+1
S48
NO
最終ストライプ番号まで完了？
YES
エンド

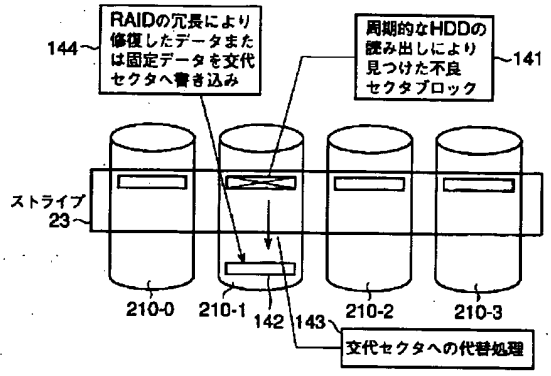
【図12】


固定データを書き込み
固定データの排他的論理和を書き込み
211
211
211
212
ストライプ
23
210-0
正常なHDD
210-1
正常なHDD
210-2
正常なHDD
210-3
修復するHDD

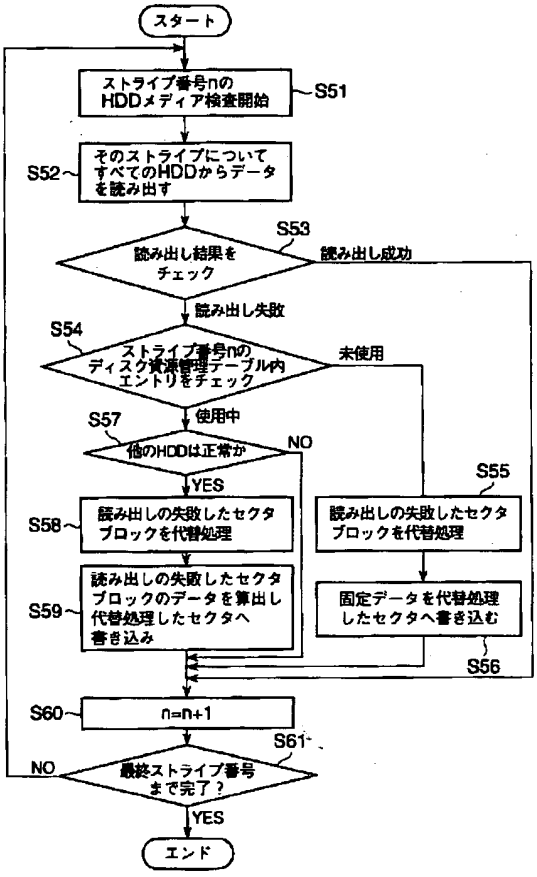
【図13】


ホスト計算機より受信した書き込みデータ
131
132
書き込み
適当な固定データ
133
134
134
書き込み
受信したデータと固定データの排他的論理和の結果
135
136
書き込み
ストライプ
23
210-0
HDD
210-1
HDD
210-2
HDD
210-3
HDD

【図14】


RAIDの冗長により修復したデータまたは固定データを交代セクタへ書き込み
144
周期的なHDDの読み出しにより見つけた不良セクタブロック
141
ストライプ23
210-0
210-1
142
143
210-2
210-3
交代セクタへの代替処理

【図15】


スタート
ストライプ番号nのHDDメディア検査開始
S51
そのストライプについてすべてのHDDからデータを読み出す
S52
読み出し結果をチェック
S53
読み出し成功
読み出し失敗
ストライプ番号nのディスク資源管理テーブル内エントリをチェック
S54
未使用
使用中
他のHDDは正常か
S57
NO
YES
S55
読み出しの失敗したセクタブロックを代替処理
S58
読み出しの失敗したセクタブロックのデータを算出し代替処理したセクタへ書き込み
S59
固定データを代替処理したセクタへ書き込む
S56
n=n+1
S60
S61
NO
最終ストライプ番号まで完了？
YES
エンド